PARAMETER ESTIMATION IN DYNAMICAL MODELS

G. EVENSEN
Nansen Environmental and Remote Sensing Center
Edvard Griegsvei 3A, N–5037 Solheimsviken, Bergen, Norway

D. DEE
General Sciences Corporation
Laurel, Maryland, U.S.A.

AND

J. SCHRÖTER
Alfred-Wegener-Institut für Polar- und Meeresforschung
Bremerhaven, Germany

Abstract. This paper will give a general introduction to the parameter estimation problem for dynamical models. The basic formulation and methodology in a parameter estimation problem will be discussed and some rather simple examples will be presented. It will be shown that even for linear dynamics the parameter estimation problem becomes nonlinear and may become extremely difficult to solve. Also, to have a well posed problem with a unique solution, care must be taken when a parameter estimation problem is formulated. The discussion leads into a conclusion that it is possible to estimate poorly known parameters in a model, at least for simple dynamical models, but care must be taken to have a consistent solution. The rule is that all paramenters which will be estimated should be added in a penalty function as weak constraints measuring their distance from a first guess in some norm. Some previous works where data assimilation methods have been used to improve estimates of poorly known model parameters or even the model bias will be briefly reviewed.

1. Introduction

Data assimilation and inverse methods are normally used for generating estimates of dynamical variables, taking into account both the information

about the dynamics from a dynamical model and the information about the true state which is contained in a set of measurements. Such techniques have also been proposed as a tool for parameter estimation in dynamical models, although only a few works in oceanography have so far considered the parameter estimation problem in this context. The idea is that it should be possible to use mathematical tools to formulate inverse problems for parameter estimation if there is additional information available about the true state of the ocean in the form of measurements. Thus, one may attempt to search for model parameters resulting in a model solution which is close to the observed state.

The parameter estimation problem has been more extensively treated in other fields and many applications have been based on the theory outlined in [10] and [11]. Most relevant are the applications in groundwater flow, [7], [8], [9].

Currently there only exist a few publications where such inverse calculations have been carried out for oceanographic applications and these have considered relatively simple dynamical models. Examples are [22], which estimated the phase speeds in a reduced-gravity model of equatorial waves, using a strong constraint variational formulation, and [24] (with a correction in [25]), which used a modified one-dimensional Ekman layer model in combination with variational optimal control techniques to simultaneously estimate the surface wind drag coefficient and the vertical profile of the eddy viscosity from observed data. This was a rather early work in this field and they had some errors in the formulation and solution method which made the problem ill-posed. In a paper [14] this problem was revisited and solved using more recent methodologies. In the paper [21], a strong constraint formulation for a quasi-geostrophic model was used to estimate the initial stream function for the model, the Rossby radius of deformation, the friction coefficient and two parameters describing the curl of the wind stress. Two other works, [23] and [26], attempt to determine the "optimal" nudging coefficients using adjoint techniques in a parameter estimation context. Some of the results from these papers will be discussed in Section 5, and have clearly given a very important contribution to our current understanding of the parameter estimation problem. A recent review, [20], discusses several previous parameter estimation applications in atmospheric and oceanic models.

One would expect that these methods would be much more widely used. However, the parameter estimation problem often becomes extremely complicated to solve and is nonlinear even if the dynamical model itself is linear and it is in many cases extremely difficult to find a proper solution. It is also a fact that most of the scientists who are working in data assimilation are still busy solving the general data assimilation problem for

realistic Ocean General Circulation Models (OGCM) and have given the parameter estimation problem less attention. However we still think the parameter estimation problem should be further explored and there is still a lot of challenging work to be done in the years to come. The major interest will probably be parameter estimation in OGCMs but possibly even more related to calibration of other types of dynamical models, e.g. marine ecosystem and sediment models.

This paper will discuss in some detail the general parameter estimation problem and how it can be formulated in a consistent manner. A general solution procedure will be presented which can be applied with a large class of dynamical models. Some of the previously published works on parameter estimation will be briefly reviewed and their methodology and results will be discussed.

2. Formulation of a parameter estimation problem

We will now look at the general formulation of the parameter estimation problem and discuss the properties of the equations that will have to be solved.

2.1. A MODEL AND OBSERVATIONS

For illustrational purposes we will use a a simple nonlinear ordinary differential equation when formulating and discussing the parameter estimation problem. The model for ψ with an initial condition is given as

$$\frac{d\psi}{dt} = f(\alpha, \psi), \tag{1}$$

$$\psi(0) = \psi_0. \tag{2}$$

Here f is a nonlinear model operator and it is assumed that α is a poorly known parameter or constant in the model. The system as it is has one unique solution for a given value of α. Thus if the value of α shall be improved we need additional information. This can be given by introducing a set of observations of the model variable taken at various locations in time (and space if the space dimensions are included). The vector of observations can be written as

$$\boldsymbol{d} = \boldsymbol{\mathcal{M}}[\psi], \tag{3}$$

where the linear operator $\boldsymbol{\mathcal{M}}$ relates the observations to the model variable.

As an example of the linear measurement operator $\boldsymbol{\mathcal{M}}[\psi]$, which relates the observations $\boldsymbol{d}$ to the state $\psi(t)$, a direct measurement would have a

measurement functional of the form

$$\mathcal{M}_i[\psi] = \int_0^T \psi(t)\delta(t - T_i)\, dt = \psi(T_i), \tag{4}$$

with T_i as the measurement location in time, and the subscript i denoting a component of the measurement functional which is a vector with dimension m equal to the number of measurements.

In general the problem becomes over determined since it will not be possible to chose one value for α that gives a model trajectory which fits the observations exactly, thus, no solution exist when the observations are introduced as additional constraints. This is resolved by allowing the observations to contain errors with some prescribed statistical properties. The measurement equation then becomes

$$\boldsymbol{d} = \boldsymbol{\mathcal{M}}[\psi^t] + \boldsymbol{\epsilon}, \tag{5}$$

where ψ^t is the unknown true state and $\boldsymbol{\epsilon}$ is a vector of measurement errors with mean $\overline{\boldsymbol{\epsilon}} = 0$ and error covariance $\overline{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}} = \boldsymbol{w}^{-1}$.

The overbar denotes some kind of averaging operator. In stochastic estimation theory this is the *expectation operator*, defined in terms of the set of all possible realizations of the stochastic-dynamic system whose state is to be estimated [19]. The precise definition of this operator requires the joint probability distribution of forecast and observation errors. In a more practical setting, it might represent an *ensemble average*, defined in terms of certain computational realizations of a deterministic model of the system (e.g. a forecast model) [18]. In many operational data assimilation systems the esemble average is simply replaced by a *time average*—that is, forecast and observation error means and covariances are computed by averaging in time the errors associated with a single realization of the system [1].

There are now two essential items that should be noted:

1. There will *always* be some information available about realistic or physically acceptable values of the poorly known parameter α. Such information could be that the value of the parameter should stay within some physically acceptable limits. This information should always be included in the inverse formulation through a first guess value and its uncertainty.
2. Unless such a first guess estimate is included in the formulation it will in many cases become practically impossible to determine how many observations are needed to find a unique solution for the parameters. Alternatively, given a particular set of observations, it can not be ensured that there exists a unique solution. This is mainly related to the fact that the information in the observations may not be independent. There may then be several values for the parameters leading to solutions which interpolate the data exactly.

Thus, the first guess value of the parameter should as a rule always be included in the inverse formulation. As will be seen later this ensures a unique solution even without any observations.

An equation for the parameter can now be defined as

$$\alpha = \alpha_0 + \tilde{\alpha}, \tag{6}$$

where α_0 is now the first or best guess value and $\tilde{\alpha}$ is the error in α_0 assumed to have statistics defined by $\overline{\tilde{\alpha}} = 0$ and $\overline{\tilde{\alpha}^2} = Q_{\tilde{\alpha}\tilde{\alpha}}$. The inverse of the covariance is defined as $W_{\tilde{\alpha}\tilde{\alpha}} = Q^{-1}_{\tilde{\alpha}\tilde{\alpha}}$.

We will now look at some different formulations of the inverse problem.

2.2. STRONG CONSTRAINT FORMULATION: ALTERNATIVE ONE

The parameter estimation problem can now be formulated by seeking a value of α which results in the model solution which is closest to the observations. Since we have introduced statistical assumptions involving only the two first moments (mean and covariance) it is natural to use a least squares estimator and we define the penalty function

$$\begin{aligned} \mathcal{J}_1[\alpha] &= \tilde{\alpha} W_{\tilde{\alpha}\tilde{\alpha}} \tilde{\alpha} + \boldsymbol{\epsilon}^{\mathrm{T}} \boldsymbol{w} \boldsymbol{\epsilon} \\ &= (\alpha - \alpha_0) W_{\tilde{\alpha}\tilde{\alpha}} (\alpha - \alpha_0) + (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi])^{\mathrm{T}} \boldsymbol{w} (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi]). \end{aligned} \tag{7}$$

Thus we attempt to search for a value of α, not too far from α_0, which gives the model solution closest to the observations. We then have to minimize the penalty function (7) with respect to α, under the constraint that the initial condition and the model equation are satisfied exactly. This is a so called constrained minimization problem which can most easily be solved by creating a Lagrangian functional where the model and the initial conditions are added as strong constraints using Lagrangian multipliers, μ and λ,

$$\mathcal{L}_1[\psi, \alpha, \mu, \lambda] = \mathcal{J}_1 + \mu(\psi - \psi_0) + \int_0^T \lambda \left(\frac{d\psi}{dt} - f(\alpha, \psi) \right) dt. \tag{8}$$

This functional can now be minimized using variational calculus. Clearly, the variation with respect to μ and λ just gives us back the forward model (1) with its initial condition (2) which must be satisfied exactly. The variation with respect to ψ results in an equation for λ with a final condition at $t = T$. This is the so called adjoint equation which can be integrated backward in time forced by the data misfits. It is given is Section 2.6 as (22) with final condition (23) where the Euler–Lagrange equations will be discussed in some more detail. The variation with respect to α results in an expression for the gradient of the Lagrangian with respect to α which can be used in a descent algorithm (see equation 24 in Section 2.6).

2.3. STRONG CONSTRAINT FORMULATION: ALTERNATIVE TWO

The formulation above does not take into account that the initial condition may contain some errors. An error in the initial condition will lead to a situation where the estimated value for α also attempts to correct for this, and a poor estimate for α may be found.

This problem can is easily resolved by allowing for errors in the initial condition on the form

$$\psi(0) = \psi_0 + a, \tag{9}$$

where the error a is assumed to satisfy the statistical assumptions $\overline{a} = 0$ and $\overline{a^2} = Q_{aa}$. Again we define the weight as the inverse of the variance $W_{aa} = Q_{aa}^{-1}$. The first guess initial condition can then be added as a weak constraint and the penalty function becomes

$$\begin{aligned} \mathcal{J}_2[\psi(0), \alpha] &= \tilde{\alpha} W_{\tilde{\alpha}\tilde{\alpha}} \tilde{\alpha} + a W_{aa} a + \boldsymbol{\epsilon}^{\mathrm{T}} \boldsymbol{w} \boldsymbol{\epsilon} \\ &= (\alpha - \alpha_0) W_{\tilde{\alpha}\tilde{\alpha}} (\alpha - \alpha_0) + (\psi - \psi_0) W_{aa} (\psi - \psi_0) \\ &+ (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi])^{\mathrm{T}} \boldsymbol{w} (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi]). \end{aligned} \tag{10}$$

Thus, a solution is searched for which is close to the observations $\boldsymbol{d}$, the initial condition ψ_0 and the first guess value α_0, while the model equation is assumed to be satisfied exactly. Again a Lagrangian is constructed to include the forward model equation

$$\mathcal{L}_2[\psi, \alpha, \lambda] = \mathcal{J}_2 + \int_0^T \lambda \left(\frac{d\psi}{dt} - f(\alpha, \psi) \right) dt. \tag{11}$$

The Euler-Lagrange equations will be similar to the ones for the Lagrangian (8) but now contain an additional term accounting for the errors in the initial condition. See Section 2.6 for a discussion of the Euler-Lagrange equations.

2.4. WEAK CONSTRAINT GENERAL INVERSE FORMULATION

Another problem with the formulations above is that they assume the model equation is perfect except for the poorly known α. Thus, it has not been taken into account that the model may be corrupted by other errors, such as those related to insufficient resolution, neglected physics, linearizations and other approximations that are commonly adapted in most dynamical models to make the equations more tractable. Thus, if there are model errors of significant magnitude compared to the real world which are observed by the measurements, a similar situation arises as for the previous case, where the estimate for the parameter α attempts to correct for other errors than those related to the particular parameterization.

There is also another issue as discussed in detail in [16] which becomes important for nonlinear models. If the model is describing unstable or chaotic dynamics such as the ocean and atmosphere, the strong constraint formulation will have a particular problem estimating the correct initial conditions. This is caused by the strong sensitivity of the model solution and then also the penalty function with respect to small changes in the initial conditions. By including the model as a weak constraint this problem is completely eliminated and we get a better posed problem to solve [16]. The reason is that the inverse estimate can deviate from the exact model trajectory and forget very past and future information (relative to the predictability time of the dynamics). Thus, after a few predictability times the estimate will not be much influenced by the inital codition if there are also sufficient observations to assimilate.

Thus, we should account for general model errors in the formulation. This can easily be done by allowing the model equation to contain errors. To sum up we now have the following system of equations defining the problem:

$$\frac{d\psi}{dt} = f(\alpha, \psi) + q, \tag{12}$$

$$\psi(0) = \psi_0 + a, \tag{13}$$

$$\alpha = \alpha_0 + \tilde{\alpha}, \tag{14}$$

$$\boldsymbol{d} = \boldsymbol{\mathcal{M}}[\psi^t] + \boldsymbol{\epsilon}. \tag{15}$$

The only new variable is $q(t)$ which represents the model errors other than those related to the value of α. Again we make a hypothesis about the model errors; $\overline{q(t)} = 0$ and $\overline{q(t_1)q(t_2)} = Q_{qq}(t_1, t_2)$. Note also that we have neglected all cross correlations between the different error terms, which in most cases is a valid assumption (the errors are normally generated by independent sources). The inverse of $Q_{qq}(t_1, t_2)$ is defined as $W_{qq}(t_1, t_2)$ through the relation

$$\int_0^T W_{qq}(t_1, t_2) Q_{qq}(t_2, t_3)\, dt_2 = \delta(t_1 - t_3). \tag{16}$$

We can now formulate the general weak constraint inverse problem for

the system (12–15) in terms of a penalty function

$$\begin{aligned}\mathcal{J}[\psi,\alpha] &= \tilde{\alpha}W_{\tilde{\alpha}\tilde{\alpha}}\tilde{\alpha} + \int_0^T\int_0^T q(t_1)W_{qq}(t_1,t_2)q(t_2)\,dt_1\,dt_2 + aW_{aa}a + \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{\epsilon}\\ &= (\alpha-\alpha_0)W_{\tilde{\alpha}\tilde{\alpha}}(\alpha-\alpha_0)\\ &+ \int_0^T\int_0^T\left\{\left(\frac{d\psi(t_1)}{dt_1} - f(\alpha,\psi(t_1))\right)\right.\\ &\qquad\left. W_{qq}(t_1,t_2)\left(\frac{d\psi(t_2)}{dt_2} - f(\alpha,\psi(t_2))\right)\right\}dt_1\,dt_2\\ &+ (\psi-\psi_0)W_{aa}(\psi-\psi_0) + (\boldsymbol{d}-\boldsymbol{\mathcal{M}}[\psi])^{\mathrm{T}}\boldsymbol{w}(\boldsymbol{d}-\boldsymbol{\mathcal{M}}[\psi]).\end{aligned} \tag{17}$$

It should now be noted that if the statistical hypothesis about normally distributed errors is correct, then the minimum of $\mathcal{J}$ (which is the optimal variance minimizing estimate) will be the maximum likelihood estimate for ψ and α. Thus, using weights which are the inverses of the error covariances is justifiable.

By minimizing this variational functional an estimate is searched for which minimizes the distance to the observations, the initial condition, the first guess value for α and at the same time "almost" satisfies the dynamical model equation.

2.5. EULER–LAGRANGE EQUATIONS

It is now instructive to present the Euler-Lagrange equations for the penalty function $\mathcal{J}$ which is derived by taking the variation of $\mathcal{J}$ with respect to ψ and α. Thus, ψ and α define an extremum of the penalty function if

$$\mathcal{J}[\psi+\delta\psi,\alpha] - \mathcal{J}[\psi,\alpha] = \mathcal{O}(\delta\psi^2), \tag{18}$$

$$\mathcal{J}[\psi,\alpha+\delta\alpha] - \mathcal{J}[\psi,\alpha] = \mathcal{O}(\delta\alpha^2), \tag{19}$$

when $\delta\psi$ and $\delta\alpha$ approach zero.

One then gets

$$\frac{d\psi}{dt} = f(\alpha,\psi) + \int_0^T Q_{qq}(t,t_1)\lambda(t_1)dt_1, \tag{20}$$

$$\psi(0) = \psi_0 + Q_{aa}\lambda(0), \tag{21}$$

$$\frac{d\lambda}{dt} = -f_\psi(\alpha,\psi)\lambda - \boldsymbol{\mathcal{M}}^{\mathrm{T}}[\delta]\boldsymbol{w}(\boldsymbol{d}-\boldsymbol{\mathcal{M}}[\psi]), \tag{22}$$

$$\lambda(T) = 0, \tag{23}$$

$$\alpha = \alpha_0 + Q_{\tilde{\alpha}\tilde{\alpha}}\int_0^T f_\alpha(\alpha,\psi)\lambda\,dt. \tag{24}$$

Since the model is nonlinear a Taylor expansion is used for $f(\alpha, \psi)$ and the first derivatives of f with respect to α and ψ are introduced as f_α and f_ψ. Note that λ is defined through the relation (20). From this general set of equations we get the Euler–Lagrange equations for the Lagrangian $\mathcal{L}_2$ by setting $Q_{qq} = 0$ and for $\mathcal{L}_1$ by setting $Q_{qq} = Q_{aa} = 0$.

The expression $\boldsymbol{\mathcal{M}}[\delta]$ is a short notation for a vector of elements

$$\mathcal{M}_i[\delta(t - t_2)] = \int_0^T \delta(t - t_2)\delta(t_2 - T_i)\, dt_2, \tag{25}$$

which are equal to 1 when the model time variable is equal to T_i and zero elsewhere.

The system of Euler–Lagrange equations defines the extrema of $\mathcal{J}$. The system consists of the original forward model forced by a term that is proportional to the adjoint variable λ in (20). The magnitude of this term is defined by the model error covariance, thus large model errors give a large contribution through the forcing term. The forward model is integrated from an initial condition which also contains a similar correction term proportional to the adjoint variable. The equation for λ can be integrated backward in time from a final condition, while forced by delta functions scaled by the residual between the measurement and forward model estimate at each measurement location. Thus, the forward model needs knowledge of the adjoint variable to be integrated, and the backward model needs the forward variable at measurement locations. We therefore have a coupled boundary value problem in time where the forward and backward models must be solved simultaneously.

2.6. MODEL BIAS

A final but not less important remark is related to a bias in the model. It is normally easier to remove an eventual bias in observations, e.g. by collecting control observations using other instruments, than the bias in the model. If initial conditions are estimated from observations they may also be unbiased. Thus the important problem will be to account for model bias which we have assumed is zero in the formulation above. Clearly, if there is a significant bias in the model, this will corrupt both the inverse problem for the model solution and the parameter estimation problem. We may end up with a totally unphysical estimate for α which attempts to correct for a model bias caused by a completely different physical approximation.

Estimation of model bias may be considered as a special case of the parameter estimation problem. If the model (12) is rewritten as

$$\frac{d\psi}{dt} = f(\alpha, \psi) + \beta + q, \tag{26}$$

β can now represent the unknown model bias which is constant in time. Thus, there is no approximation related to setting $\overline{q(t)} = 0$. An equation is needed for β,

$$\beta = \beta_0 + \tilde{\beta}. \tag{27}$$

It is assumed that β_0 is a first guess value for the bias which would typically be zero in many applications. The error $\tilde{\beta}$ in the first guess for the bias is assumed to be normally distributed with $\overline{\tilde{\beta}} = 0$ and $\overline{\tilde{\beta}^2} = Q_{\tilde{\beta}\tilde{\beta}}$. The inverse of the covariance is defined as $W_{\tilde{\beta}\tilde{\beta}} = Q^{-1}_{\tilde{\beta}\tilde{\beta}}$. A weak constraint variational formulation for this problem is

$$\begin{aligned}
\mathcal{J}[\psi,\alpha] &= \tilde{\alpha} W_{\tilde{\alpha}\tilde{\alpha}} \tilde{\alpha} + \tilde{\beta} W_{\tilde{\beta}\tilde{\beta}} \tilde{\beta} \\
&\quad + \int_0^T \int_0^T q(t_1) W_{qq}(t_1,t_2) q(t_2)\, dt_1\, dt_2 + a W_{aa} a + \boldsymbol{\epsilon}^{\mathrm{T}} \boldsymbol{w} \boldsymbol{\epsilon} \\
&= (\alpha - \alpha_0) W_{\tilde{\alpha}\tilde{\alpha}} (\alpha - \alpha_0) + (\beta - \beta_0) W_{\tilde{\beta}\tilde{\beta}} (\beta - \beta_0) \\
&\quad + \int_0^T \int_0^T \left\{ \left(\frac{d\psi(t_1)}{dt_1} - f(\alpha, \psi(t_1)) - \beta \right) \right. \\
&\qquad\qquad \left. W_{qq}(t_1,t_2) \left(\frac{d\psi(t_2)}{dt_2} - f(\alpha, \psi(t_2)) - \beta \right) \right\} dt_1\, dt_2 \\
&\quad + (\psi - \psi_0) W_{aa} (\psi - \psi_0) + (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi])^{\mathrm{T}} \boldsymbol{w} (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi]).
\end{aligned} \tag{28}$$

The Euler–Lagrange equations for this problem becomes

$$\frac{d\psi}{dt} = f(\alpha,\psi) + \beta + \int_0^T Q_{qq}(t,t_1) \lambda(t_1) dt_1, \tag{29}$$

$$\psi(0) = \psi_0 + Q_{aa} \lambda(0), \tag{30}$$

$$\frac{d\lambda}{dt} = -f_\psi(\alpha,\psi)\lambda - \boldsymbol{\mathcal{M}}[\delta] \boldsymbol{w} (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi]), \tag{31}$$

$$\lambda(T) = 0, \tag{32}$$

$$\alpha = \alpha_0 + Q_{\tilde{\alpha}\tilde{\alpha}} \int_0^T f_\alpha(\alpha,\psi)\lambda\, dt, \tag{33}$$

$$\beta = \beta_0 + Q_{\tilde{\beta}\tilde{\beta}} \int_0^T \lambda\, dt. \tag{34}$$

Thus, we are now estimating both the value for α and the bias β in the model. In Section 4 an alternative method will be illustrated where the model bias is corrected using a sequential data assimilation algorithm.

3. A solution method for the parameter estimation problem.

At this stage it should be pointed out that the inverse problem may become highly nonlinear both because of the nonlinearities in the model, but also by the introduction of the parameter as an additional variable in the system. It may in many cases be impossible to find a solution using traditional methods involving the Euler–Lagrange equations. However, a methodology will now be presented which has proven successful with linear and weakly nonlinear models.

3.1. ITERATION OF PARAMETERS

The system of Euler–Lagrange equations is strongly coupled through the parameters and the dependent variable ψ and it is also nonlinear. To simplify the system it is possible to define an iteration for the unknown parameters. Note that the equations for α and β, (33) and (34), are the gradient of the penalty function with respect to the parameters. By defining the iterations for the parameters as

$$\alpha^{k+1} = \alpha^k - \gamma_\alpha \left(\alpha^k - \alpha_0 - Q_{\tilde{\alpha}\tilde{\alpha}} \int_0^T f_\alpha(\alpha^k, \psi^k) \lambda^k dt \right), \tag{35}$$

$$\beta^{k+1} = \beta^k - \gamma_\beta \left(\beta^k - \beta_0 - Q_{\tilde{\beta}\tilde{\beta}} \int_0^T \lambda^k dt \right), \tag{36}$$

the problem is now reduced to solving a sequence of inverse problem for ψ and λ for each iterate k of the parameters α and β. The two factors γ_α and γ_β are just the step lengths in the gradient descent iterations. The new estimates for ψ and λ are then used to iterate the parameters through (35) and (36). Thus, if these iterations converge, the problem is now reduced to solving the following equations for each iteration (iteration superscripts have been dropped):

$$\frac{d\psi}{dt} = f(\alpha, \psi) + \beta + \int_0^T Q_{qq}(t, t_1) \lambda(t_1) dt_1, \tag{37}$$

$$\psi(0) = \psi_0 + Q_{aa} \lambda(0), \tag{38}$$

$$\frac{d\lambda}{dt} = -f_\psi(\alpha, \psi)\lambda - \boldsymbol{\mathcal{M}}^{\mathrm{T}}[\delta] \boldsymbol{w} (\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi]) \tag{39}$$

$$\lambda(T) = 0. \tag{40}$$

These are the Euler–Lagrange equations for the standard inverse or data assimilation problem without any parameters to estimate.

3.2. SOLUTION OF THE "ALTERNATIVE ONE" PROBLEM

The problem defined by $\mathcal{L}_1$ can now be solved easily since both Q_{aa} and Q_{qq} are zero. This removes the coupling between the forward and backward equations in (37–40), and given an estimate for the parameters one can always find ψ from one forward integration, and λ from a following backward integration. It is suggested that this approach is used to test if the iterations of the parameters converge before one of the more sophisticated methods which are discussed below is implemented.

3.3. SOLUTION OF THE "ALTERNATIVE TWO" PROBLEM

The assumption that the model is perfect, $Q_{qq} = 0$, apparently simplifies the Euler–Lagrange equations (37–40). A data assimilation or inverse method which is very commonly used for this special case is the so called "adjoint method". The method iterates the initial condition in the same way as has been done for the parameters above using that the equation (38) is just the gradient of $\mathcal{L}_2$ with respect to the initial conditions. We can define the iteration

$$\psi^{k+1} = \psi^k - \gamma_\psi(\psi^k - \psi_0 - Q_{aa}\lambda^k), \tag{41}$$

where again γ_ψ is a step length in the gradient descent method.

Thus, it is now possible to solve the parameter estimation problem by iterating the equations (35), (36) and (41). Each iteration provides an estimate for the two parameters and the initial condition which makes it possible to solve for the current iterate of ψ and λ, needed to update the parameters and initial condition again. Note, that various strategies should to be tried such as using several iterations for the initial condition to have convergence for ψ and λ before updating the parameters again. This may be especially important for high dimensional problems.

Note that this method searches for the solution in a space with dimension equal to the number of state variables in the model. For many realistic ocean models this dimension becomes extremely large, $\mathcal{O}(10^{5-8})$.

3.4. REPRESENTER SOLUTION FOR A LINEAR INVERSE PROBLEM

The representer method is an elegant methodology which can solve the linear inverse problem without any iterations and where the solution is searched for in a space with dimension equal to the number of measurements.

Assume now that the model dynamics are linear. The cases with weakly nonlinear and strongly nonlinear dynamics will be treated next. The model operator $f(\alpha, \psi)$ is now for illustrational purposes written $f\alpha\psi$ where f is

now a scalar for this particular model, $\alpha \simeq 1$ can be interpreted as a poorly known correction to f. For this linear model we now get the Euler–Lagrange equations of the form (where the parameters are now assumed given)

$$\frac{d\psi}{dt} = f\alpha\psi + \beta + \int_0^T Q_{qq}(t,t_1)\lambda(t_1)dt_1, \tag{42}$$

$$\psi(0) = \psi_0 + Q_{aa}\lambda(0), \tag{43}$$

$$\frac{d\lambda}{dt} = -f\alpha\lambda - \boldsymbol{\mathcal{M}}[\delta]^{\mathrm{T}}\boldsymbol{w}(\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi]), \tag{44}$$

$$\lambda(T) = 0. \tag{45}$$

Start by assuming that a solution can be written as

$$\psi(t) = \psi_F(t) + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{r}(t), \tag{46}$$

$$\lambda(t) = \lambda_F(t) + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{s}(t), \tag{47}$$

where the dimensions of the vectors are all equal to the number of measurements, m. Assuming this form for the solution is equivalent to saying that the minimizing solution is a first guess model solution plus a linear combination of time dependent influence functions or representers $\boldsymbol{r}(t)$, one for each measurement. For a comprehensive discussion of this method see [2].

Inserting (46) and (47) into the Euler–Lagrange equations (42–45) and choosing first guesses ψ_F and λ_F which satisfy unforced exact equations

$$\frac{d\psi_F}{dt} = f\alpha\psi_F + \beta + \int_0^T Q_{qq}\lambda_F, \tag{48}$$

$$\psi_F(0) = \psi_0, \tag{49}$$

$$\frac{d\lambda_F}{dt} = -f\alpha\lambda_F, \tag{50}$$

$$\lambda_F(T) = 0, \tag{51}$$

gives us the following system of equations for the vector of representers $\boldsymbol{r}$ and corresponding adjoints $\boldsymbol{s}$:

$$\boldsymbol{b}^{\mathrm{T}}\left(\frac{d\boldsymbol{r}}{dt} - f\alpha\boldsymbol{r} - \int_0^T Q_{qq}(t,t_1)\boldsymbol{s}(t_1)dt_1\right) = 0, \tag{52}$$

$$\boldsymbol{b}^{\mathrm{T}}\left(\boldsymbol{r}(0) - Q_{aa}\boldsymbol{s}\right) = 0, \tag{53}$$

$$\boldsymbol{b}^{\mathrm{T}}\left(\frac{d\boldsymbol{s}}{dt} + f\alpha\boldsymbol{s}\right) + \boldsymbol{\mathcal{M}}^{\mathrm{T}}[\delta]\boldsymbol{w}\left(\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi_F + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{r}]\right) = 0, \tag{54}$$

$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{s}(T) = 0. \tag{55}$$

By defining the so far undetermined vector of coefficients $\boldsymbol{b}$ to be given by

$$\boldsymbol{b} = \boldsymbol{w}\left(\boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi_F + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{r}]\right), \tag{56}$$

equation (54) now becomes

$$\boldsymbol{b}^{\mathrm{T}}\left(\frac{d\boldsymbol{s}}{dt} + \boldsymbol{s} + \boldsymbol{\mathcal{M}}[\delta]\right) = 0, \tag{57}$$

which now decouples the forward and backward Euler–Lagrange equations.

Using that $\boldsymbol{b}$ in general is nonzero we now have the following set of equations in addition to (48–51): first equations (52) and (53) becomes

$$\frac{d\boldsymbol{r}}{dt} = f\alpha\boldsymbol{r} + \int_0^T Q_{qq}\boldsymbol{s}\,dt, \tag{58}$$

$$\boldsymbol{r}(0) = Q_{aa}\boldsymbol{s}, \tag{59}$$

for the representers, and from equations (57) and (55) we get

$$\frac{d\boldsymbol{s}}{dt} = -f\alpha\boldsymbol{s} - \boldsymbol{\mathcal{M}}[\delta], \tag{60}$$

$$\boldsymbol{s}(T) = 0, \tag{61}$$

for the "adjoints" of the representers.

From equation (56) one easily gets the linear system

$$(\boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}] + \boldsymbol{w}^{-1})\boldsymbol{b} = \boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi_F], \tag{62}$$

which can be solved as soon as the representer matrix $\boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}]$ has been generated.

The equations (60) and (61) for $\boldsymbol{s}$ can now be solved as a sequence of final value problems since they are decoupled from the forward equations (58) and (59) for the representers. As soon as $\boldsymbol{s}$ is found the representers can be solved for. This provides the information needed for solving the system (62) for $\boldsymbol{b}$ and the final estimate is found by solving the Euler–Lagrange equation of the form

$$\frac{d\psi}{dt} = f\alpha\psi + \beta + \int_0^T Q_{qq}\lambda\,dt, \tag{63}$$

$$\psi(0) = \psi_0 + Q_{aa}\lambda(0), \tag{64}$$

$$\frac{d\lambda}{dt} = -f\alpha\lambda - \boldsymbol{\mathcal{M}}^{\mathrm{T}}[\delta]\boldsymbol{b}, \tag{65}$$

$$\lambda(T) = 0. \tag{66}$$

The numerical load is $2m+3$ model integrations, with m the number of measurements, but note that only 2 states need to be stored in space and time. If the solution is constructed directly from (46) all the representers need to be stored.

In a parameter estimation problem for a linear model the procedure is as follows:

1. Start with $\alpha^0 = \alpha_0$ and $\beta^0 = \beta_0$.
2. Solve the current iterate of the Euler-Lagrange equations (42–45) for ψ and λ using the representer method.
3. Use ψ and λ in the iterations (35) and (36) to find new estimates for the parameters.
4. Go to 2 and repeat until convergence.

The methodology outlined in this Section was used to estimate a vertical diffusion parameter and the wind-drag coefficient in an one-dimensional Ekman model in [14].

Note that this approach can be used also for the "alternative two" problem, and in the Appendix a recently developed approach for speeding up the representer method is presented. This new approach completely avoids the calculation of the representers. The low numerical cost for this algorithm actually makes the representer method a very attractive approach for solving inverse problems.

3.5. REPRESENTER SOLUTION FOR A WEAKLY NONLINEAR INVERSE PROBLEM

The representer method can not be used directly to solve nonlinear inverse problems. However, if one can define a convergent sequence of linear iterates of the nonlinear model, the inverse problem for each linear iterate can be solved using the representer method. As an example consider the equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \cdots. \tag{67}$$

If the solution of this equation can be found from the iteration

$$\frac{\partial u_k}{\partial t} + u_{k-1}\frac{\partial u_k}{\partial x} = \cdots, \tag{68}$$

then one can also define a convergent sequence of linear inverse problems which can be solved exactly using representer expansions. In the papers [6], [4], and [3], this approach was used. The major conclusion from these papers is that the representer method can be used with realistic atmospheric prediction models to an affordable numerical cost and that the iteration on the Euler–Lagrange equation does converge quickly for the particular

models used. The representer method can be characterized as the optimal solution method for linear and weakly nonlinear inverse problems, both in terms of the problem minimized (full weak constraint problem) and in terms of numerical cost.

3.6. DISCUSSION

Note that the equation for $\boldsymbol{b}$, (62), is similar to the one solved in the analysis scheme in the standard Kalman filter [15]. Actually the only difference is that here we have a covariance matrix with corresponding representers or influence functions in space and time, while in the Kalman filter the time dimension is excluded. Thus, the solution is sought for in a similar form. In the Kalman filter the representer or influence functions are defined as the measurements of the error covariance matrix ($\boldsymbol{MP}$), with $\boldsymbol{M}$ the measurement matrix and $\boldsymbol{P}$ the error covariance matrix. A similar interpretation can be used here if we define the error covariance $C(t_1, t_2)$ for the first guess solution which results in an expression for the representers

$$\boldsymbol{r} = \boldsymbol{\mathcal{M}}[C]. \tag{69}$$

It can be shown that the representers are the measurements of the reproducing kernel of the inner product defined by the penalty function, and the reproducing kernel is also the error covariance of the first guess solution [2].

It can also be shown that if we seek a solution of the form

$$\psi(t) = \psi_F(t) + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{r}(t) + g(t), \tag{70}$$

where $g(t)$ is an arbitrary function orthogonal to the space spanned by the representers, then the minimizing solution must have $g(t) \equiv 0$, [2].

This also shows that the solution is sought for in the m-dimensional space spanned by the representers. Thus, we have reduced the infinite dimensional problem defined by the penalty function to an m-dimensional problem.

3.7. STRONGLY NONLINEAR INVERSE PROBLEMS

In highly nonlinear cases one may obtain a solution using a brute force direct substitution method such as the gradient descent method used for the Lorenz model in [16]. If the penalty function $\mathcal{J}$ is discretized in time (and space if the spatial dimensions are included), it is relatively easy to derive an analytical expression for the gradient of $\mathcal{J}$ with respect to the full model state $\boldsymbol{\psi}$ and the parameter α, and a minimum can be found using a descent algorithm. It may of course be a local minimum if the penalty function becomes nonlinear enough and in that case one may get

a step further using statistical methods such as Hybrid Monte Carlo and annealing [17].

4. Model bias correction in a sequential data assimilation system

Here we briefly review a different approach for dealing with model bias that was recently developed [13]. They considered the effect of model bias on state estimates produced by a sequential data assimilation algorithm, and showed how such an algorithm can be modified to properly account for this effect. Their approach was motivated by the fact that the majority of current operational data assimilation systems used for weather prediction and climate studies involves sequential estimation algorithms. Typically in atmospheric four-dimensional data assimilation a short-term (e.g., 6h) forecast is used as a background estimate for the true atmospheric state, and this background estimate is then combined with new observations to produce a so-called analysis. The analysis is subsequently used to initialize the forecast model, and the process is repeated. These assimilation systems are extremely large and complex, involving more than 10^6 state variables, and very little is known about the error characteristics of the forecast model. In this application it is therefore of great practical importance that any algorithm for bias correction is not highly sensitive to assumptions about the nature of the bias, and that its implementation does not require a complete redesign of the assimilation system.

The main idea behind the sequential approach is that one can estimate the time-mean forecast error directly from the observations, assuming that the observations themselves are unbiased. This estimate can then be used to correct the forecast prior to producing the analysis. Thus, rather than estimating a set of model bias parameters, the algorithm estimates and corrects for the effect of the model bias upon the short-term forecast. As it turns out, sequential estimation of the time-mean forecast error can be accomplished by means of existing components of a sequential data assimilation system. It is this feature in particular that makes this algorithm interesting from a practical point of view.

4.1. NOTATION

Let the n-vector $\boldsymbol{\psi}_k^f$ denote a model forecast valid for time t_k, and $\boldsymbol{\psi}_k^t$ the unknown true state at that time. The *forecast error* is then simply

$$\boldsymbol{e}_k^f \equiv \boldsymbol{\psi}_k^f - \boldsymbol{\psi}_k^t. \tag{71}$$

For a m_k-vector $\boldsymbol{d}_k$ of measurements generated by a particular instrument at time t_k, the *observation error* is

$$\boldsymbol{\epsilon}_k \equiv \boldsymbol{d}_k - \boldsymbol{M}_k \boldsymbol{\psi}_k^t. \tag{72}$$

The linear m_k-vector function $\boldsymbol{M}_k$ is the *discrete forward observation operator*, mapping model variables to the data type associated with the instrument.

We use the following notation for the forecast error means and covariances:

$$\mathbf{b}_k^f \equiv \overline{\boldsymbol{e}_k^f}, \qquad \boldsymbol{P}_k^f \equiv \overline{(\boldsymbol{e}_k^f - \mathbf{b}_k^f)(\boldsymbol{e}_k^f - \mathbf{b}_k^f)^{\mathrm{T}}}. \tag{73}$$

We use the roman letter $\mathbf{b}$ here to distinguish forecast bias from model bias parameters $\boldsymbol{\beta}$; generally $\mathbf{b}_k^f = \mathbf{b}_k^f(\boldsymbol{\beta})$.

We assume that observations are unbiased, and

$$\boldsymbol{R}_k \equiv \overline{\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^{\mathrm{T}}}. \tag{74}$$

We do not consider possible covariances between forecast and observation errors here.

We may use the observation operator in (72) to compare forecasts with observations. The *observed-minus-forecast residuals* defined by

$$\boldsymbol{v}_k = \boldsymbol{d}_k - \boldsymbol{M}_k \boldsymbol{\psi}_k^f \tag{75}$$

contain information about the actual observation and forecast errors. This is clearly expressed by

$$\boldsymbol{v}_k = \boldsymbol{\epsilon}_k - \boldsymbol{M}_k \boldsymbol{e}_k^f, \tag{76}$$

which obtains from (71) and (72).

The residual means and covariances are easily obtained from (76):

$$\overline{\boldsymbol{v}_k} \approx -\boldsymbol{M}_k \mathbf{b}_k^f, \tag{77}$$

$$\overline{(\boldsymbol{v}_k - \overline{\boldsymbol{v}_k})(\boldsymbol{v}_k - \overline{\boldsymbol{v}_k})^{\mathrm{T}}} \approx \boldsymbol{R}_k + \boldsymbol{M}_k \boldsymbol{P}_k^f \boldsymbol{M}_k^{\mathrm{T}}. \tag{78}$$

4.2. FORECAST BIAS ESTIMATION

We can rewrite (76) as

$$\boldsymbol{v}_k = -\boldsymbol{M}_k \mathbf{b}_k^f + \boldsymbol{\eta}_k, \tag{79}$$

where $\boldsymbol{\eta}_k$ is a random vector with first and second moments

$$\overline{\boldsymbol{\eta}_k} \approx 0, \tag{80}$$

$$\overline{\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\mathrm{T}}} \approx \boldsymbol{R}_k + \boldsymbol{M}_k \boldsymbol{P}_k^f \boldsymbol{M}_k^{\mathrm{T}}. \tag{81}$$

This follows from (76) and (78) by noting that $\boldsymbol{\eta}_k = \boldsymbol{v}_k - \overline{\boldsymbol{v}_k}$.

Equation (79) can be regarded as a *measurement model* for the forecast bias $\mathbf{b}_k^f$. It expresses the relationship between the observations, the forecast, and the forecast bias. For each fixed time t_k, the dimension m_k of $\boldsymbol{v}_k$ is generally less than the dimension n of $\mathbf{b}_k^f$, so that (79) cannot be inverted without additional information.

This is a familiar problem: incomplete observations of an unknown quantity must be supplemented with information from a *state model* in order to compensate for the lack of data. When the unknown is the state of a natural system (e.g., the Earth's atmosphere), some kind of predictive evolution model may be used for this purpose (e.g., a general circulation model). In case the unknown quantity is forecast bias, however, it is not immediately obvious what form the state model should take.

Generally the model should express any known or assumed spatial, temporal, and multivariate relationships satisfied by the bias field. For example, suppose that we wish to estimate the time-mean forecast error, averaged over a time period exceeding synoptic time scales. A reasonable state model for $\mathbf{b}_k^f$ is then simply *persistence*, i.e.,

$$\mathbf{b}_k^f = \mathbf{b}_{k-1}^f. \tag{82}$$

From (79), (80), and (82) we easily obtain a recursive estimate for the forecast bias $\mathbf{b}_k^f$:

$$\widehat{\mathbf{b}}_k^f = \widehat{\mathbf{b}}_{k-1}^f - \boldsymbol{L}_k \left[\boldsymbol{v}_k + \boldsymbol{M}_k \widehat{\mathbf{b}}_{k-1}^f \right] \tag{83}$$

where $\boldsymbol{L}_k$ is—for now—an arbitrary matrix. When initialized with an *a priori* estimate $\widehat{\mathbf{b}}_0^f$, equation (83) is a sequential algorithm for estimating the forecast bias associated with a data assimilation system. The algorithm uses the observed-minus-forecast residuals $\boldsymbol{\epsilon}_k$ produced by the system.

We can show [13] that the optimal choice of $\boldsymbol{L}_k$ is

$$\boldsymbol{L}_k = \boldsymbol{P}_{k-1}^b \boldsymbol{M}_k^{\mathrm{T}} \left[\boldsymbol{M}_k \boldsymbol{P}_k^f \boldsymbol{M}_k^{\mathrm{T}} + \boldsymbol{R}_k \right]^{-1} \tag{84}$$

with $\boldsymbol{P}_k^b$ the error covariance of the bias estimate $\widehat{\mathbf{b}}_k^f$:

$$\boldsymbol{P}_k^b \equiv \overline{(\widehat{\mathbf{b}}_k^f - \mathbf{b}_k^f)(\widehat{\mathbf{b}}_k^f - \mathbf{b}_k^f)^{\mathrm{T}}}. \tag{85}$$

It is possible to derive a recursion for $\boldsymbol{P}_k^b$ as well; combined with (84) this yields the Kalman filter for the system (79, 82). It is more practical, however, to specify the covariance matrix $\boldsymbol{P}_k^b$ based on simplifying assumptions.

This is the usual procedure for obtaining the forecast and observation error covariances $\boldsymbol{P}_k^f$ and $\boldsymbol{R}_k$ for operational data assimilation systems.

An example of a class of bias estimation error covariance specifications is simply

$$\boldsymbol{P}_k^b = \gamma \boldsymbol{P}_k^f \tag{86}$$

where γ is a scalar. Assuming that the forecast error covariances themselves are modeled in terms of a set of parameters $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_N)$ [12], a more general class of specifications is

$$\boldsymbol{P}_k^b(\boldsymbol{\alpha}) = \gamma \boldsymbol{P}_k^f(\boldsymbol{\alpha}). \tag{87}$$

The parameter vector $\boldsymbol{\alpha}$ may include, for example, variance parameters, spatial decorrelation length scales, etc.

4.3. ON-LINE BIAS CORRECTION

We can apply the algorithm (83)–(84) to estimate forecast bias associated with any existing data assimilation system, using stored observed-minus-forecast residuals. Moreover, if the system involves sequential analysis updates, forecast bias can be estimated on the fly. We can then use the estimates to correct the biased forecasts on-line and to produce unbiased analyses. The complete algorithm is:

forecast: (88)

$$\boldsymbol{\psi}_k^f = \boldsymbol{F}_k \widetilde{\boldsymbol{\psi}}_{k-1}^a \tag{89}$$

$$\widetilde{\boldsymbol{\psi}}_k^f = \boldsymbol{\psi}_k^f - \widehat{\mathbf{b}}_{k-1}^f \tag{90}$$

bias update: (91)

$$\widehat{\mathbf{b}}_k^f = \widehat{\mathbf{b}}_{k-1}^f - \boldsymbol{L}_k \left[\boldsymbol{d}_k - \boldsymbol{M}_k \widetilde{\boldsymbol{\psi}}_k^f \right] \tag{92}$$

$$\boldsymbol{L}_k = \boldsymbol{P}_k^b \boldsymbol{M}_k^{\mathrm{T}} \left[\boldsymbol{M}_k \boldsymbol{P}_k^f \boldsymbol{M}_k^{\mathrm{T}} + \boldsymbol{R}_k \right]^{-1} \tag{93}$$

$$\widetilde{\widetilde{\boldsymbol{\psi}}}_k^f = \boldsymbol{\psi}_k^f - \widehat{\mathbf{b}}_k^f \tag{94}$$

analysis: (95)

$$\widetilde{\boldsymbol{\psi}}_k^a = \widetilde{\widetilde{\boldsymbol{\psi}}}_k^f + \boldsymbol{K}_k \left[\boldsymbol{d}_k - \boldsymbol{M}_k \widetilde{\widetilde{\boldsymbol{\psi}}}_k^f \right] \tag{96}$$

$$\boldsymbol{K}_k = (\boldsymbol{P}_k^f - \boldsymbol{P}_k^b) \boldsymbol{M}_k^{\mathrm{T}} \left[\boldsymbol{M}_k (\boldsymbol{P}_k^f - \boldsymbol{P}_k^b) \boldsymbol{M}_k^{\mathrm{T}} + \boldsymbol{R}_k \right]^{-1} \tag{97}$$

The operator $\boldsymbol{F}_k$ represents the forecast model; $\widetilde{\boldsymbol{\psi}}_k^f$ and $\widetilde{\widetilde{\boldsymbol{\psi}}}_k^f$ are the *a priori* and *a posteriori* bias-corrected forecasts, respectively, and $\widetilde{\boldsymbol{\psi}}_k^a$ is the unbiased analysis. Note that for $\boldsymbol{P}_k^b = 0$ and with $\widehat{\mathbf{b}}_0^f \equiv 0$, the algorithm reduces to the familiar sequential analysis update equations.

5. Some previous parameter estimation works

The paper [21] pointed at several important issues regarding the formulation of the parameter estimation inverse problem. They used a quasi-geostrophic ocean model for the Gulf Stream extension and assimilated sea surface heights from Geosat radar altimeter data. A strong constraint formulation was used where the penalty function included the data misfit and a Laplacian smoothing term plus a quasi geostrophic model as strong constraints, and it was minimized using the adjoint method. The control variables were the initial stream function for the model, the Rossby radius of deformation, the friction coefficient and two parameters describing the wind-stress curl.

From the discussion in the previous sections this appears to be an "ill posed" problem since none of the first guesses are penalized. Without the smoothness constraint the problem was not well posed and it was possible to fit the observations closely and get a penalty function smaller than the noise level. This also lead to an estimate for the initial stream function which was very noisy and physically unacceptable. Furthermore, it could not be ensured that this was the only initial field fitting the observations. Thus, the problem was likely to be underdetermined.

By adding a Laplacian smoothness constraint on the initial conditions the problem becomes a "semi-norm". Thus the situation improves considerably and a smooth estimate for the initial condition is found. However, it can still not be proved that this solution is unique, not even for linear dynamics. For example, assume that no observations are available. In this case there are infinitively many solutions. If a single observation is included, there will still be infinitively many solutions. One could continue adding observations but it is impossible to know how many are needed to close the problem. One could be lead to believe that if the number of observations is larger than the number of control variables this would be sufficient. However, in general this is not the case. The reason is that the observations will normally not be independent. For example, two observations may be located close together (say within one Rossby deformation length). They will then provide some dependent information.

A very important result from this study was that when using the smoothness constraint it was impossible to estimate the diffusion parameter. The

problem is that with a smooth initial condition the parameter attempts to become negative which leads to an exponential growth of the noise that remains in the model solution. Clearly, the reason for this is that the noisier solution will have a better chance to get close to the noisy observations than a smooth model solution.

What has been learned from this work is that it is important to penalize the deviation from the first guesses of all control variables. This will ensure that the estimated amplitudes stay within a physically acceptable range. Furthermore, a regularization constraint is needed for all control variables. Here a smoothness constraint was used for the initial conditions. Alternatively such a constraint can be incorporated by using nondiagonal weights in the penalty function.

In the works [24], [25], some similar problems where encountered. In [24] a penalty function was defined where the first guesses of the diffusion coefficient and the wind drag coefficient were penalized in addition to the residual between the model results and the observations. However, there was no penalty on the first-guess initial conditions. Without such a penalty, every choice of initial conditions can be used, and there may exist many initial conditions resulting in a solution which interpolates the data and gives a penalty function equal to zero as shown in [5]. [25] included such a term although the resulting "gradient" equation for the initial condition was incorrect. With a correct equation for the initial condition the only missing issue would be the regularization of the vertical diffusion parameter. This is needed to ensure that a smooth estimate is found for the diffusion parameter.

The paper [14] reformulated this inverse problem using a weak constraint formulation with proper penalties on the first guesses and regularizations of all control parameters. The Euler–Lagrange equations were derived and solved by the methodology outlined in the previous sections using iterations on the parameters and the representer method.

There was a substantial difference in the values for the diffusion coefficient and the wind drag obtained in the two works. It should be noted that [24] *replaced* the first-guess values of the parameters with the *current estimate* in each iteration of the parameters. These should of course be kept constant. Clearly, [24] solved a different inverse problem in each iteration and did not have any real penalty of the first guesses at all. Actually, it is not clear from their figures that the iterations did converge.

6. Summary

A general introduction has been given to the formulation and solution of the parameter estimation problem for dynamical models. A methodology has

been presented where the dynamical model with its initial condition and poorly known parameters are all treated as weak constraints in a variational inverse formulation. The variational formulation penalizes deviations from the exact model equation and distances from the first guesses for the control variables. The minimum of the variational functional is an estimate which almost satisfies the model equations and at the same time is "close" to the observations and the first guesses of the initial condition and parameters. From this general weak constraint formulation it is easy to obtain the Euler–Lagrange equations for suboptimal or less general strong constraint formulations. The basic method proposed here for estimating poorly known parameters defines a gradient descent iteration for the parameters, while the remaining inverse problem (with the parameters assumed given) is solved using the representer method [2]. The method should be applicable for many problems as long as they are not too strongly nonlinear. For an extensive discussion of an application of the method, see [14], where the vertical diffusion and the wind-drag was estimated using the method described here.

A. Speeding up the representer method

With a large number of observations the standard formulation of the representer method becomes very expensive to compute. However there are ways to get around this problem as discussed in [3]. First make the following observations:

1. Only $\boldsymbol{b}$ is needed to compute the estimate from (63–66).
2. If a conjugate gradient method is used to solve (62), then the representer matrix $\boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}]$ itself is not needed, only the representer matrix times a vector must be computed in each iteration.

Point 2 above can be illustrated by the following: rewrite first the linear system (62) for the coefficients $\boldsymbol{b}$, as

$$\boldsymbol{A}\boldsymbol{b} = \boldsymbol{h}, \tag{98}$$

with $\boldsymbol{A} = \boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}] + \boldsymbol{w}^{-1}$ and $\boldsymbol{h} = \boldsymbol{d} - \boldsymbol{\mathcal{M}}[\psi_F]$. Then we define the quadratic functional

$$\mathcal{J}_{\boldsymbol{b}}[\boldsymbol{b}] = \frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{b} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h}, \tag{99}$$

where the gradient with respect to $\boldsymbol{b}$ becomes

$$\nabla_{\boldsymbol{b}}\mathcal{J}_{\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{b} - \boldsymbol{h}. \tag{100}$$

Thus, the minimum of $\mathcal{J}_{\boldsymbol{b}}$ is also where the gradient $\boldsymbol{A}\boldsymbol{b} - \boldsymbol{h} = 0$. A gradient descent iteration can now be defined as

$$\boldsymbol{b}^{k+1} = \boldsymbol{b}^{k} - \gamma_{\boldsymbol{b}}\left(\boldsymbol{A}\boldsymbol{b}^{k} - \boldsymbol{h}\right) \tag{101}$$

where $\gamma_{\boldsymbol{b}}$ is just a factor used in the gradient descent algorithm. Thus, the system (98) can be solved using the iteration given by (101), where only matrix-vector multiplications need to be calculated. In practical settings a more sophisticated iteration should be used such conjugate gradient methods.

Starting by measuring (46), we get

$$\boldsymbol{\mathcal{M}}[\psi(t)] = \boldsymbol{\mathcal{M}}[\psi_F(t)] + \boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}(t)]\boldsymbol{b}. \tag{102}$$

If we assume a problem with the first guess solution $\psi_F \equiv 0$, (obtained by setting the initial condition equal to zero), we get

$$\boldsymbol{\mathcal{M}}[\psi(t)] = \boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}(t)]\boldsymbol{b}, \tag{103}$$

and thus, measuring the minimizing solution is equivalent to computing the representer matrix times $\boldsymbol{b}$. Now, define an arbitrary estimate $\tilde{\psi}$ which is obtained from the Euler–Lagrange equations with $\psi_0 = 0$, $\beta = 0$ and by replacing $\boldsymbol{b}$ with $\boldsymbol{b}^k$, i.e., $\tilde{\psi}$ is the solution of the system

$$\frac{d\tilde{\psi}}{dt} - \tilde{\psi} = Q_{qq}\lambda, \tag{104}$$

$$\tilde{\psi}(0) = Q_{aa}\lambda(0), \tag{105}$$

$$\frac{d\lambda}{dt} + \lambda = -\boldsymbol{\mathcal{M}}^{\mathrm{T}}[\delta]\boldsymbol{b}^k, \tag{106}$$

$$\lambda(T) = 0, \tag{107}$$

obtained from (63–66). Then, measuring $\tilde{\psi}$ will result in a vector equal to $\boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}(t)]\boldsymbol{b}^k$. Thus, in each iteration of the conjugate gradient algorithm, the system above is solved to calculate $\tilde{\psi}$ from a given $\boldsymbol{b}^k$. Then $\tilde{\psi}$ is measured to generate $\boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}(t)]\boldsymbol{b}^k$, and hence $(\boldsymbol{\mathcal{M}}^{\mathrm{T}}[\boldsymbol{r}(t)] + \boldsymbol{w}^{-1})\boldsymbol{b}^k$ which is needed in the algorithm.

The numerical cost for finding the minimizing solution is now equal to $2j + 3$ model integrations, where j is the required number of iterations in the conjugate gradient algorithm to solve the system of linear equations (62). Note that the "real" dimension of this system is equal to the number of independent degrees of freedom in the observations which in many cases is order of magnitudes less than the state variables to be estimated. Preconditioning can be used to speed up the conjugate gradient algorithm, i.e., if an approximation to the matrix of the linear system is available this will normally lead to a significant reduction in the number of iterations. In [3] an approach was used where an approximation to the representer matrix was calculated on coarse grid from $2m$ integrations which can be performed relatively quickly.

ACKNOWLEDGEMENT

This work has been supported by the EC–MAST–III BASYS project (MAS3–CT96–0058).

References

1. P. Bartello and H. L. Mitchell. A continuous three-dimensional model of short-range forecast error covariances. *Tellus, Ser. A*, 44:217–235, 1992.
2. A. F. Bennett. *Inverse Methods in Physical Oceanography.* Cambridge University Press, 1992.
3. A. F. Bennett, B. S. Chua, and L. M. Leslie. Generalized inversion of a global numerical weather prediction model. *Meteorol. Atmos. Phys.*, 60:165–178, 1996.
4. A. F. Bennett, L. M. Leslie, C. R. Hagelberg, and P. E. Powers. Tropical cyclone prediction using a barotropic model initialized by a generalized inverse method. *Mon. Weather Rev.*, 121:1714–1729, 1993.
5. A. F. Bennett and R. N. Miller. Weighting initial conditions in variational assimilation schemes. *Mon. Weather Rev.*, 119:1098–1102, 1990.
6. A. F. Bennett and M. Thorburn. The generalized inverse of a nonlinear quasi-geostrophic ocean circulation model. *J. Phys. Oceanogr.*, 22:213–230, 1992.
7. J. Carrera and S. Neuman. Estimation of aquifer parameters under transient and steady state conditions: 1. maximum likelihood method incorporating prior information. *Water Resour. Res.*, 22:199–210, 1986.
8. J. Carrera and S. Neuman. Estimation of aquifer parameters under transient and steady state conditions: 2. uniqueness, stability, and solution algorithms. *Water Resour. Res.*, 22:211–227, 1986.
9. J. Carrera and S. Neuman. Estimation of aquifer parameters under transient and steady state conditions: 3. application to synthetic and field data. *Water Resour. Res.*, 22:228–242, 1986.
10. G. Chavent. Identification of functional parameters in partial differential equations. In R. E. Goodson and M. Polis, editors, *Identification of Parameters in Distributed Systems*, pages 31–48. American Society of Mechanical Engineers, New York, 1974.
11. G. Chavent. Identification of distributed parameter system: About the output least square method, its implementation, and identifiability. In R. Isermann, editor, *Identification and System Parameter Estimation, Proceedings of the Fifth IFAC Symposium, Darmstadt, Federal Republic of Germany, 24–28 September 1979*, volume 1, pages 85–97, 1979.
12. D. P. Dee. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Weather Rev.*, 123:1128–1145, 1995.
13. D. P. Dee and A. M. da Silva. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, 1998. In press.
14. M. Eknes and G. Evensen. Parameter estimation solving a weak constraint variational formulation for an Ekman model. *J. Geophys. Res.*, 102(C6):12,479–12,491, 1997.
15. G. Evensen. Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. *J. Geophys. Res.*, 97(C11):17,905–17,924, 1992.
16. G. Evensen. Advanced data assimilation for strongly nonlinear dynamics. *Mon. Weather Rev.*, 125:1342–1354, 1997.
17. G. Evensen and N. Fario. A weak constraint variational inverse for the Lorenz equations using substitution methods. *J. Meteor. Soc. Japan*, 75(1B):229–243, 1997.
18. P. L. Houtekamer, L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell. A system simulation approach to ensembles prediction. *Mon. Weather Rev.*, 124:1225–1242, 1996.
19. A. H. Jazwinski. *Stochastic Processes and Filtering Theory.* Academic, San Diego,

Calif., 1970.

20. I. M. Navon. Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. *Dyn. Atmos. Oceans*, 27:55–79, 1997.
21. J. Schröter, U. Seiler, and M. Wenzel. Variational assimilation of Geosat data into an eddy-resolving model of the Gulf stream extension area. *J. Phys. Oceanogr.*, 23:925–953, 1993.
22. O. M. Smedstad and J. J. O'Brien. Variational data assimilation and parameter estimation in an equatorial Pacific Ocean model. *Prog. Oceanogr.*, 26:179–241, 1991.
23. D. R. Stauffer and J.-W. Bao. Optimal determination of nudging coefficients using the adjoint equations. *Tellus, Ser. A*, 45:358–369, 1993.
24. L. Yu and J. J. O'Brien. Variational estimation of the wind stress drag coefficient and the oceanic eddy viscosity profile. *J. Phys. Oceanogr.*, 21:709–719, 1991.
25. L. Yu and J. J. O'Brien. On the initial condition in parameter estimation. *J. Phys. Oceanogr.*, 22:1361–1364, 1992.
26. X. Zou, I. M. Navon, and F. X. L. Dimet. An optimal nudging data assimilation scheme using parameter estimation. *Q. J. R. Meteorol. Soc.*, 118:1163–1168, 1992.